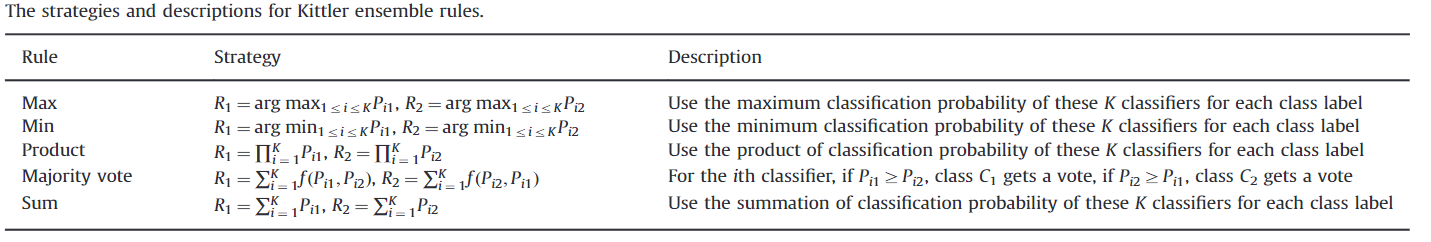
**Summary**

The following table shows the summary of the most common ensemble techniques

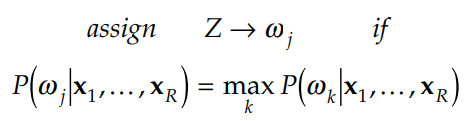


The detail explanation of the Bayesian and the product rule are presented below.

1. The Bayesian decision rule

Consider a pattern recognition problem where pattern Z is to be assigned to one of the m possible classes (). Let us assume that we have R classifiers each representing the given pattern by a distinct measurement vector. Denote the measurement vector used by the ith classifier by xi. In the measurement space each class is modeled by the probability density function and its a priori probability of occurrence is denoted . We shall consider the models to be mutually exclusive which means that only one model can be associated with each pattern.

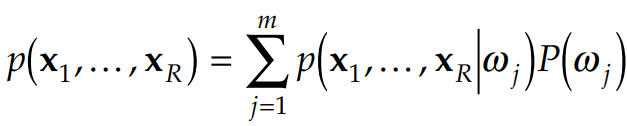
Now, according to the Bayesian theory, given measurements , the pattern, Z, should be assigned to class provided the a posteriori probability of that interpretation is maximum, i.e.

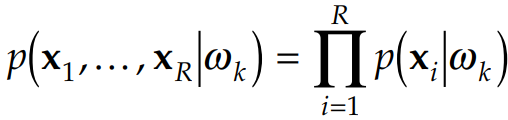
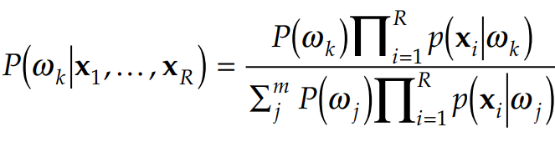


The Bayesian decision rule states that in order to utilize all the available information correctly to reach a decision, it is essential to compute the probabilities of the various hypotheses by considering all the measurements simultaneously. This is, of course, a correct statement of the classification problem but it may not be a practicable proposition. The computation of the a posteriori probability functions would depend on the knowledge of high-order measurement statistics described in terms of joint probability density functions which would be difficult to infer. We shall therefore attempt to simplify the above rule and express it in terms of decision support computations performed by the individual classifiers, each exploiting only the information conveyed by vector xi. We shall see that this will not only make rule computationally manageable, but also it will lead to combination rules which are commonly used in practice. Moreover, this approach will provide a scope for the development of a range of efficient classifier combination strategies.

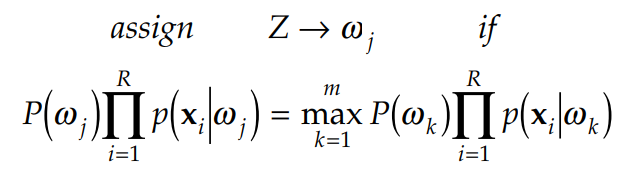
Let us rewrite the a posteriori probability using the Bayes theorem. We have

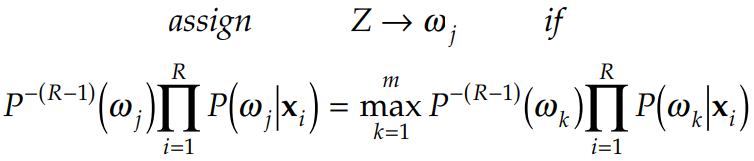
where is the unconditional measurement joint probability density. The latter can be expressed in terms of the conditional measurement distributions as



1. Product Rule

represents the joint probability distribution of the measurements extracted by the classifiers. Let us assume that the representations used are conditionally statistically independent. The use of different representations may be a probable cause of such independence in special cases. We will investigate the consequences of this assumption and write

we obtain the decision rule as follows:

or in terms of the a posteriori probabilities yielded by the respective classifiers

Refrences

1. Kittler, Josef, et al. "On combining classifiers." IEEE transactions on pattern analysis and machine intelligence 20.3 (1998): 226-239.
2. Nahin, Paul J., and John L. Pokoski. "NCTR plus sensor fusion equals IFFN or can two plus two equal five?." IEEE Transactions on Aerospace and Electronic Systems 3 (1980): 320-337.
3. Sun, Zhongbin, et al. "A novel ensemble method for classifying imbalanced data." Pattern Recognition 48.5 (2015): 1623-1637.

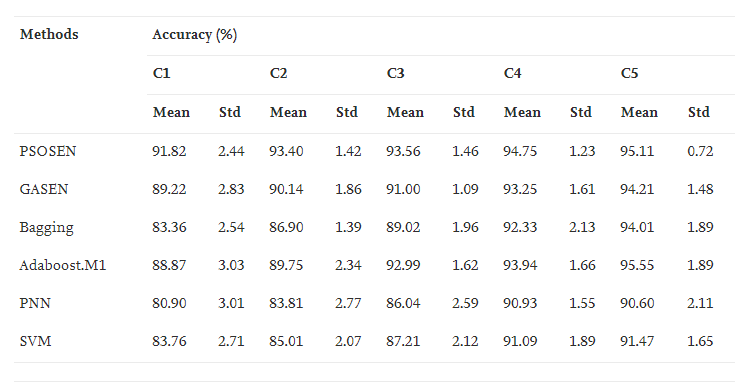
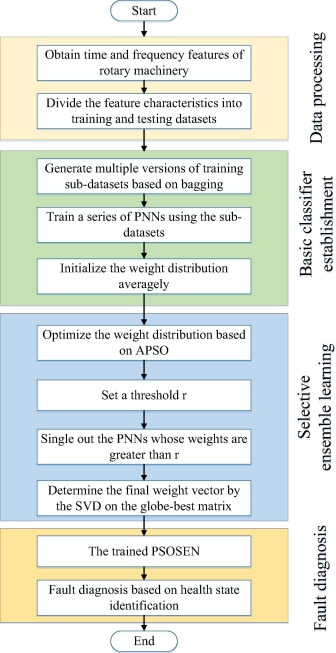
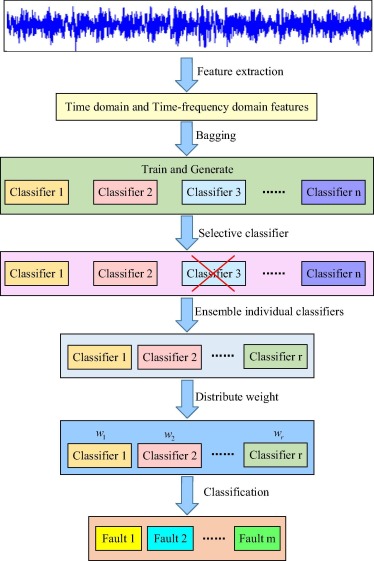
**Paper review**

1. **Fault diagnosis for rotary machinery with selective ensemble neural networks**

[Wang, Zhen-Ya, Chen Lu, and Bo Zhou. "Fault diagnosis for rotary machinery with selective ensemble neural networks." Mechanical Systems and Signal Processing 113 (2018): 112-130.]

With a goal to obtain a better generalization ability of fault diagnosis along with multiple monitored variables with corresponding fault patterns for rotary machinery systems, a novel fault diagnosis method (**particle swarm optimization based selective ensemble learning, PSOSEN**) that utilizes ensemble learning with differentiated **probabilistic neural networks (PNNs)** is proposed, where nonlinear decreasing inertia weight based **adaptive particle swarm optimization (APSO)** is employed to effectively reinforce the learning process by selecting superior individuals for integration instead of all. First, statistical features in the time domain and frequency domain are extracted and integrated from vibration signals, and feature selection based on bagging feature representation is applied to generate desirable PNNs. Second, APSO is used to improve the performance by balancing diversity and accuracy, aiming to eliminate similar individuals via weight assignation and retain the classifiers with better performance in the initial iteration. The globe-best vectors are then, by means of linear transformation, mapped into a matching matrix in which row vectors indicate the corresponding weights of the selected classifiers. Singular value decomposition (SVD) is employed on the established matrix, where an optimal weight vector is thus obtained according to the orthogonal matrices parameters. The fault diagnosis result is finally achieved by ensemble computing of PNNs based on the calculated weight coefficients.

After basic training, APSO was conducted for individual selection, where the randomly initialized weights were iterated and updated based on the fitness function, and classifiers whose weights were over the pre-set threshold were selected for final ensemble.



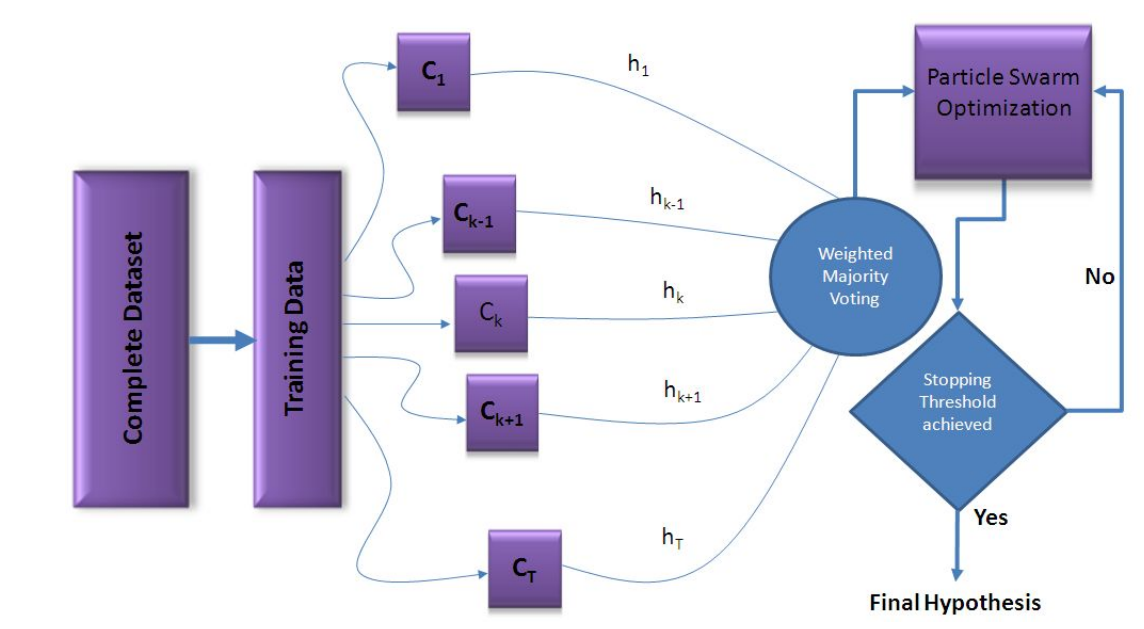
A new version of the PSOSEN based selective ensemble learning method is presented in this paper to provide value insights into intelligent fault diagnosis. The diagnosis results demonstrate that the PSOSEN model could achieve desirable accuracies and robustness than many other methods when faced with the input signals containing environmental noise and working condition fluctuations. Compared to the traditional ensemble learning methods, the proposed method performed better with far less component neural networks, which shows that the PSOSEN method is superior in classification tasks with respect to stronger generalization ability and learning efficiency. Besides, as a kind of data-driven fault identification method, it appears that the PSOSEN could have desirable applicability with respect to other rotary machineries as well.

* 1. **Optimization of Ensemble based Decision using PSO**

[Kausar, Asma, et al. "Optimization of ensemble based decision using PSO." Proceedings of the World Congress on Engineering. Vol. 2010. 2010.]

In this paper they have proposed an idea of Particle Swarm Optimization (PSO) in order to optimize weights to evaluate the competence of an expert. Weighted Majority Voting (WMV) is the most popular technique used to combine such opinions in an ensemble based classification. The weights associated to each base classifier in WMV on the basis of its competence are optimized under the influence of the basic idea of PSO. PSO has shown the stable performance on the selected datasets.

An ensemble based system is more reliable than individual classifiers when we come across the multi classification of nonlinear and complex datasets. The performance of ensemble based system can further be improved by using Practical Swarm Optimization..



1. **Multiple Classiﬁer Integration for the Prediction of Protein Structural Classes**

[Chen, Lei, et al. "Multiple classifier integration for the prediction of protein structural classes." Journal of Computational Chemistry 30.14 (2009): 2248-2254.]

Supervised classiﬁers, such as artiﬁcial neural networks, partition trees, and support vector machines, are often used for the prediction and analysis of biological data. However, choosing an appropriate classiﬁer is not straightforward because each classiﬁer has its strengths and weaknesses, and each biological dataset has its characteristics. By integrating many classiﬁers, people can avoid the dilemma of choosing an individualclassiﬁer out of many to achieve optimized classiﬁcation results.

They claimed the results of the ensembled method are better than any single machine learning algorithm collected in Weka when the same data are used. Furthermore, they introduced an integration strategy that takes care of both classiﬁer weightings and classiﬁerredundancy.

A feature selection strategy, called minimum redundancy maximum relevance (mRMR), is transferred into algorithm selection to deal with classiﬁer redundancy in this research, and the weightings are based on the performance of each classiﬁer. The best classiﬁcation results are obtained when 11 algorithms are selected by mRMRmethod, and integrated through majority votes with weightings. As a result, the prediction correct rates are68.56% and 69.29% for the basic training dataset and the independent test dataset, respectively.

**Simple Majority Voting System (SMVS)** simply counts the votes for each data. And the class, gaining the majority votes, is assigned to be the class of the data. It is a simple and popular ensemble approach since it does not require any prior knowledge or any additional complex computation for decisions.

**Weighted Majority Voting System**, In SMVS, all classiﬁers are equal, i.e., the weight of each classiﬁer is identical. However, some classiﬁers perform much better than others. Hence, assigning a higher weight to classiﬁerswith better performance should help to improve the prediction accuracy. The correct prediction rate, evaluated by ﬁvefold cross-validation on the training dataset, is taken as the weight of the classiﬁer.

**Minimum Redundancy Maximum Relevance** (mRMR)

If some classiﬁers are from the same classiﬁer family, they will have a similar classiﬁcation mechanism and more likely they will give similar classiﬁcation results. If the space partition of a classiﬁer can be simulated by other classiﬁers, this classiﬁer is redundant if other classiﬁers are used. The more redundant a classiﬁer, the worse chance it will get selected. Moreover, classiﬁers that have a higher correct prediction rate should have a higher priority to be chosen. The correct prediction rate can be reﬂected by the relevance between the results of the classiﬁers and the true categories of the data. mRMR method takes consideration of both relevance and redundancy and is adopted for the classiﬁer selection task in this article. mRMR tries to add one algorithm at a time into the algorithm list. In each round, an algorithm with maximum relevance and minimum redundancy is selected. Thus an algorithm list with the selection order can be obtained.

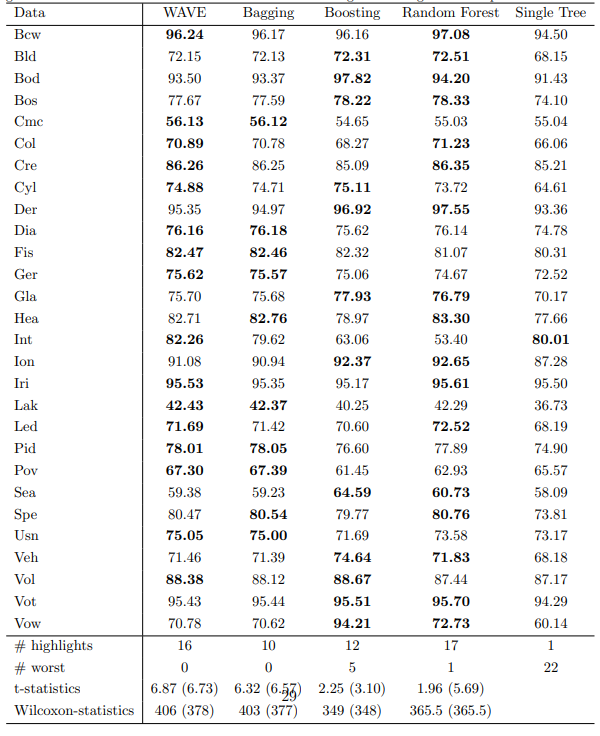
1. **A Weight-Adjusted Voting Algorithm for Ensemble of Classifiers**

[Kima, Hyunjoong, et al. "A weight-Adjusted voting algorithm for ensemble of classifiers." Journal of the Korean Statistical Society 40 (2011): 437-449.]

The paper presented a weighted voting classification ensemble method, called WAVE, that uses two weight vectors: a weight vector of classifiers and a weight vector of instances. The instance weight vector assigns higher weights to observations that are hard to classify. The weight vector of classifiers puts larger weights on classifiers that perform better on hard-to-classify instances.

One weight vector is designed to be calculated in conjunction with the other through an iterative procedure. That is, the instances of higher weights play more important role in determining the weights of classifiers, and vice versa.

The final prediction of the ensemble is obtained by the voting using the optimal weight vector of classifiers. To compare the performance between a simple majority voting and the proposed weighted voting, both of the voting methods to bootstrap aggregation and investigated the performance on 28 data sets. The result shows that the proposed weighted voting performs significantly better than the simple majority voting in general.



Remarks: the performance diffrence presented in this paper has a marginal in contrast with the other algorithms presented here.

1. **Classifier Ensemble Selection Using Genetic Algorithm for Named Entity Recognition**

[Ekbal, Asif, and Sriparna Saha. "Classifier ensemble selection using genetic algorithm for named entity recognition." Research on Language and Computation 8.1 (2010): 73-99.]

A classifier ensemble technique based on genetic algorithm (GA) for named entity recognition (NER) is proposed in this study. They assumed that the classifiers based on different feature representations can be effectively combined together using GA to achieve better performance. The proposed approach is also able to find the appropriate ensemble approach, i.e. either majority voting or weighted voting.

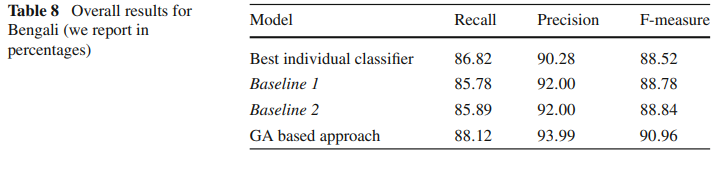
Maximum entropy (ME) model is used as a base to generate a number of different classifiers depending upon the various representations of the available features.

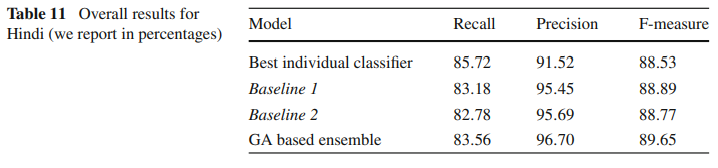
Based on the experimental resutlts, they claimed that the GA based ensemble attains the performance which is superior to all the individual classifiers as well as the conventional baseline ensembles.

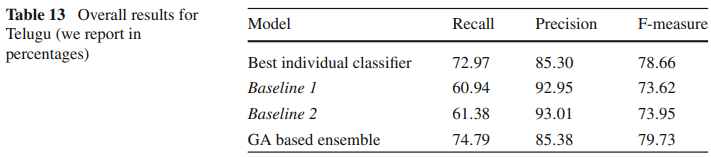
Here, GA is used to search for the most suitable combinations of the existing classifiers. Another important issue of classifier ensemble is to find out the appropriate method of combining the outputs of the classifiers. There are two types of methods, namely majority voting and weighted voting for combining the classifiers. In general, weighted voting is found to perform better than majority voting, but there are some situations where majority voting could perform superior. Thus, it is crucial to determine the appropriate way of classifier combination, i.e. either majority voting or weighted voting. In this paper, they also showed the effectiveness of GA to determine this.

They used maximum entropy (ME) as a base classifier. Depending on the various available feature combinations, different versions of this classifier are made. All the classifiers make use of language independent features that include various contextual and orthographic word-level information.

The following are sample results from the paper for three languages under test







1. **Application of Classifier Integration Model with Confusion Table to Audio Data Classification**

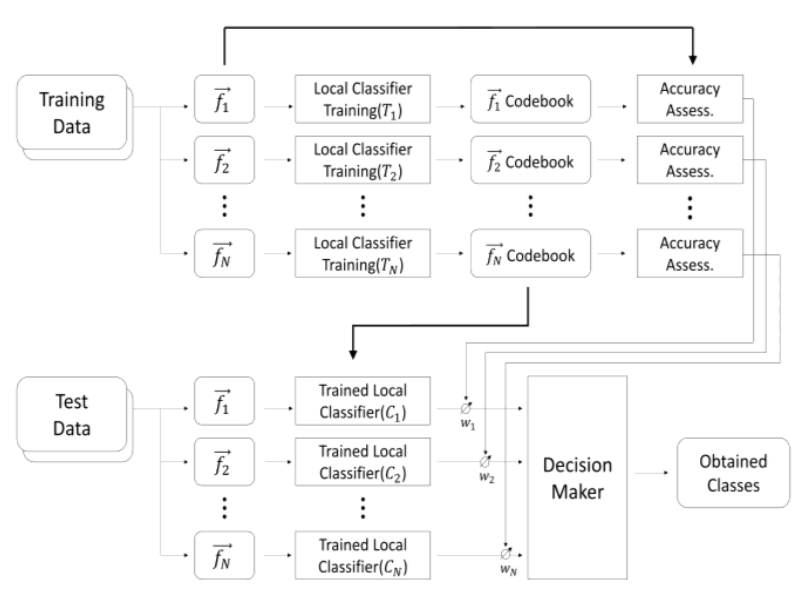
[Jang, Miso, and Dong-Chul Park. "Application of Classifier Integration Model with Confusion Table to Audio Data Classification." International Journal of Machine Learning and Computing 9.3 (2019).]

Satellite Image Classification Using a Classifier Integration Model

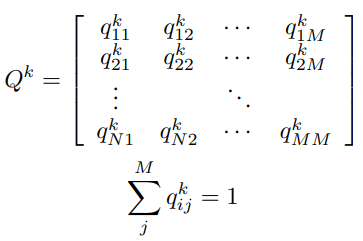
[Park, Dong-Chul, et al. "Satellite image classification using a classifier integration model." 2011 9th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA). IEEE, 2011.]

In order to enhance the classification accuracy for CIM, this paper addresses an efficient method to combine the outputs of all the local classifiers. While CIM uses only the diagonal elements of each confusion table of the local classifier, the CIM with Confusion Table(CIM-CT) method is designed to use the non-diagonal elements in addition to the diagonal elements of confusion tables.

This model was first proposed to cope with the problems experienced when training data with various features. The figure below shows a schematic diagram of CIM. The CIM can utilize all the available N features extracted from data by utilizing a number of local classifiers k, 1 ≤ k ≤ 𝑁, while conventional classifiers use only selected features at once by concatenating these features.



the CIM considers how each local classifier with different feature vector as its input performs on each class whether it assesses correct class or not. The accuracy information, or tendency, of each local classifier during the training procedure is recorded as a form of the confusion table in CIM. The confusion table for the local classifier k can be formulated as shown



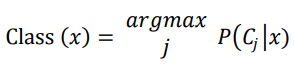
where represents the probability that the classifier k, classifies the data as Class j when the data is from of Class i

and M denotes the number of classes.

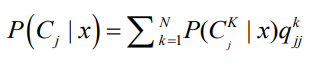
in addition to the probability that each specific local classifier made correct classifications in the past, the tendency how each specific local classifier made incorrect classification can give us valuable information when the assessment of class for a given data is made. This idea of utilizing the tendency for each local classifier to make misclassifications can be formulated as follows:



The assessment of class for a given data x can be made as follows:



The performance of this method is compared with the the orgignal methodd which only utilizes the diagonal elements, which is defined by



In the two papers they used audio data and satelige images for their experiment and they claimed their method works better. The follwing is sample result from the audio dataset experiment